HoloYolo: A proof-of-concept study for marker-less surgical navigation of spinal rod implants with augmented reality and on-device machine learning

von Atzigen, Marco ; Liebmann, Florentin ; Hoch, Armando ; Bauer, David E ; Snedeker, Jess Gerrit ; Farshad, Mazda ; Fürnstahl, Philipp

Abstract: BACKGROUND Existing surgical navigation approaches of the rod bending procedure in spinal fusion rely on optical tracking systems that determine the location of placed pedicle screws using a hand-held marker. METHODS We propose a novel, marker-less surgical navigation proof-of-concept to bending rod implants. Our method combines augmented reality with on-device machine learning to generate and display a virtual template of the optimal rod shape without touching the instrumented anatomy. Performance was evaluated on lumbosacral spine phantoms against a pointer-based navigation benchmark approach and ground truth data obtained from computed tomography. RESULTS Our method achieved a mean error of 1.83 ± 1.10 mm compared to 1.87 ± 1.31 mm measured in the marker-based approach, while only requiring 21.33 ± 8.80 s as opposed to 36.65 ± 7.49 s attained by the pointer-based method. CONCLUSION Our results suggests that the combination of augmented reality and machine learning has the potential to replace conventional pointer-based navigation in the future.

DOI: https://doi.org/10.1002/rcs.2184

Marco von Atzigen (Orcid ID: 0000-0002-6441-5905)

HoloYolo: A proof-of-concept study for marker-less surgical navigation of spinal rod implants with augmented reality and on-device machine learning

Marco von Atzigen*[,1,2], Florentin Liebmann*[,1,2], Armando Hoch[1,3], David E. Bauer[3], Jess Gerrit Snedeker[2,3], Mazda Farshad[3], Philipp Fürnstahl[1]

[1] Research in Orthopedic Computer Science, Balgrist University Hospital, University of Zurich, Zurich, Switzerland
[2] Laboratory for Orthopaedic Biomechanics, ETH Zurich, Zurich, Switzerland
[3] Orthopaedic Department, Balgrist University Hospital, University of Zurich, Zurich, Switzerland
* These authors contributed equally

Marco von Atzigen
Balgrist University Hospital
Forchstrasse 340
8008 Zurich
Switzerland
Tel.: +41 44 510 70 31
ORCID: 0000-0002-6441-5905
marco.vonatzigen@balgrist.ch

Financial support: This work is part of "SURGENT" under the umbrella of University Medicine Zurich/Hochschulmedizin Zürich, Switzerland. The NVIDIA Quadro P6000 used for this research was donated by the NVIDIA Corporation.

Category: Original Article

Word count: 5881
Number of Figures: 7
Number of Tables: 0

1 Abstract

Background: Existing surgical navigation approaches of the rod bending procedure in spinal fusion rely on optical tracking systems that determine the location of placed pedicle screws using a hand-held marker.
Methods: We propose a novel, marker-less surgical navigation proof-of-concept to bending rod implants. Our method combines augmented reality with on-device machine learning to generate and display a virtual template of the optimal rod shape without touching the instrumented anatomy. Performance was evaluated on lumbosacral spine phantoms against a pointer-based navigation benchmark approach and ground truth data obtained from Computed Tomography.
Results: Our method achieved a mean error of 1.83±1.10 mm compared to 1.87±1.31 mm measured in the marker-based approach, while only requiring 21.33±8.80 s as opposed to 36.65±7.49 s attained by the pointer-based method.
Conclusion: Our results suggests that the combination of augmented reality and machine learning has the potential to replace conventional pointer-based navigation in the future.

Keywords: Machine Learning · Augmented Reality · Object Detection · Rod Bending · Spinal Fusion · Surgical Navigation

2 Introduction

A constant aging of the society and advances in surgical techniques have led to a rapid increase of spine surgeries within the last decade [1]. Among the different surgical approaches, spinal fusion is one of the most frequently performed procedures. In the United States alone, 200 000 spinal fusion surgeries are annually performed [2] and roughly 60% of the procedures performed in our institution are fusion surgeries. Spinal fusion is indicated when patients are suffering from lower back pain due to instability, scoliosis, disk degeneration [3] or when previous conservative treatment was unsuccessful [4–6].

The instrumentation of a spinal fusion begins with the bilateral placement of screws in the pedicles of each affected vertebra. A precise placement of the pedicle screws is crucial as deviations from the targeted trajectory can result in injuries of the spinal cord, nerve roots or blood vessels. After successful screw placement, a rod implant is adopted to the patient anatomy such that it rigidly connects the vertebrae via the pedicle screws (see Figure 1). Manual bending of the rod implant is a tedious and time-consuming process. The surgeon has to move back and forth between the surgical site and the bending bench until the shape and length of the rod implant matching the instrumented anatomy has been found.

Figure 1: Instrumented anatomy: The bent rod is connected with the pedicle screws to fuse the vertebrae.

The challenging anatomy and the risk of causing serious or even fatal injuries make spinal surgery a high-risk intervention. Various surgical navigation solutions have been established to reduce the risk of failures and increase surgical precision. Pedicle screw placement is the most frequently navigated step in the instrumentation of the spine. Navigation is performed either by patient-specific instruments [7], optical tracking systems [8,9] or, more recently, by augmented reality based solutions [10,11]. Several approaches for the navigation of bending the rod implant were also introduced [12,13]. The commercially available Bendini rod bending system (NuVasive, Inc., San Diego, CA, USA), for example, captures the 3D positions of implanted pedicle screw heads by means of a pointing device equipped with infrared-reflective spheres and an optical infrared tracking system [12]. From these 3D positions, the rod geometry can be calculated and transferred to the surgeon as a set of rod bending instructions. A more sophisticated approach was recently presented by Wanivenhaus et al. [13] who made use of augmented reality to display a virtual template of the target rod using the HoloLens (Microsoft, Redmond, WA, USA). Alike the Bendini system, they also used a pointing device with an image-based marker to estimate the 3D positions of the pedicle screw heads. Although the navigation of rod bending demonstrated clinical and biomechanical benefits over the conventional method [12,13], it has not yet established in spinal surgery.

We believe that the need of using special pointing devices with attached optical markers is outdated and one of the main reasons for the low clinical acceptance of these approaches. Acquiring screw positions manually bears the risk of a misfitting rod [13]. Pointing devices and markers also cause considerable costs and efforts, as they are sterilizable medical devices that come into contact with patient anatomy. In practice many surgeons do not like markers at all because they complicate the surgical workflow.

In this study, we would like to present a purely machine learning driven navigation proof-of-concept for rod bending using the 1st generation HoloLens. The key idea is to use the live streams of the stereo cameras to detect the implanted screw heads and reconstruct their 3D poses. The problem of object detection from videos has been extensively studied for decades in the field of computer vision and hence a variety of algorithms are available that could potentially solve the task of screw head detection. Examples of recent general-purpose

convolutional neural networks (CNN) for object detection are Faster R-CNN [14], Single Shot Multibox Detector (SSD) [15] or YOLO [16–18]. Machine learning based object detection approaches have already been employed in a medical scope. In laparoscopic or robotic surgery, for example, surgical tool detection is achieved either by means of image-based classification [19], tool segmentation [20–22], recurrent networks [23] or by estimating the pose of tools and structures for monocular [24–27] and stereo input [28–30] video streams.

Compared to endoscopic interventions, open surgeries pose other challenges such as decreased object visibility due to the depth of the wound, dynamically changing appearance of implants due to partial blood coverage and difficult lighting conditions in the OR. Additionally, due to the demands on computational power and software architecture, these methods do not qualify for running locally on the HoloLens.

In this paper, we propose a method to reconstruct 3D pedicle screw head positions which could be used for marker-less augmented reality based surgical navigation of rod implants. After display calibration, our approach works out-of-the-box without the need of anatomy registration or any server infrastructure to stream data. We introduce HoloYolo as an efficient CNN-based object detection and position estimation method for the HoloLens. With our approach, the 3D positions of the implanted screw heads can be detected at interactive rates by combining stereo reconstruction with clustering for outlier removal. The precision and time effort of our method was evaluated on two lumbosacral spine phantoms and compared to a benchmark approach of Wanivenhaus et al. [13] and ground-truth data obtained from Computed Tomography (CT) scans.

3 Methods

Our method consists of four parts. In the first part a CNN is used to constantly determine the 2D positions of all visible screw head centers in the grayscale video streams of the left and right forward-facing environmental cameras (section 3.1). In a second step, correspondences between the detected centers in the left and right images are obtained (section 3.2). Then, we utilize the found correspondences to triangulate the 3D positions of the screw heads (section 3.3). Lastly, the candidate screw centers are clustered into a set of final screw head positions as described in section 3.4. The process is repeated until all screw head centers have been determined and the surgeon accepts the calculated centers upon visual inspection. The overall pipeline is illustrated in Figure 2.

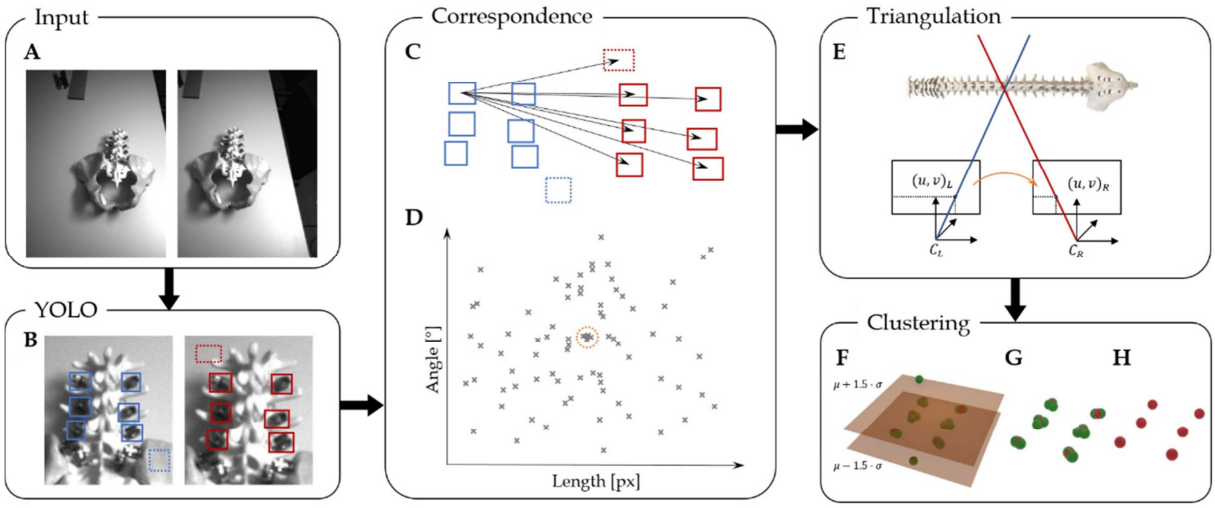


Figure 2: Our proposed pipeline. ***Input (A):*** *Stereo input images of the two front-facing environmental cameras of the HoloLens. Images are grayscale at a resolution of 480×640 px.* ***Detection (B):*** *HoloYolo extracts the bounding boxes around the pedicle screw heads. Dashed boxes denote false detections. Note that the images are magnified for visualization purposes.* ***Correspondence Finding (C and D):*** *A vector is drawn from each detection in the left image to each detection in the right image (only shown for one detection). Vectors are transformed into polar coordinates where similar vectors are represented by accumulations of points. Each point in the denoted accumulation (orange) is associated to a correspondence.* ***Triangulation (E):*** *Given the correspondences, the 3D screw head locations can be determined by means of triangulation.* ***Clustering (F-H):*** *All found points over all past frames are cumulated and outliers are removed (F). The resulting points are subject to k-means++ clustering (G). The found cluster centers are the final screw head positions (H).*

3.1 Estimating the 2D positions of the screw heads

Employing a CNN on the Microsoft HoloLens poses multiple challenges. One limitation is the operating system of HoloLens which only supports Universal Windows Platform (UWP) [31] applications and hence existing TensorFlow or PyTorch implementations of CNNs cannot be used in a straight-forward way. However, the main difficulty is the very limited computational power of the HoloLens which makes the use of larger state-of-the-art CNNs hardly possible. For instance, performing a bounding box inference on a single image takes roughly 150 times longer on the HoloLens than on the NVIDIA Quadro P6000 GPU. We have evaluated different object detection methods and found approaches that regress a bounding box for each detected object as the best trade-off between accuracy and performance. The idea is to use the center of the bounding box as an estimate of the center of the screw head. We decided to base our approach on Tiny YOLOv2 [17], which is a light-weight CNN for bounding box regression. The network was converted into a single class detector by adapting the number of filters in the last convolutional layer such that the output shape conforms to the required single class output.

In UWP applications, Microsoft supports the use of Windows Machine Learning (WinML) API to employ artificial neural networks (ANN). Existing ANN can be integrated in WinML using the Open Neural Network Exchange (ONNX) format. Tiny YOLOv2 is implemented in a C-based deep learning framework called darknet. There is no direct native conversion from darknet to ONNX, whereas both TensorFlow and PyTorch models support the transfer to ONNX. To this end, a PyTorch implementation of Tiny YOLOv2[1] was adapted to import custom-trained models from darknet. We remained using the darknet

[1] https://github.com/purelyvivid/yolo2_onnx

implementation for training due to the possibility of maintaining higher training speeds. The resulting PyTorch model was then converted to its final ONNX representation using a dedicated PyTorch package.

In order to train our network, we constructed a data set consisting of 500 labelled RGB-images of a lumbosacral spine phantom with ten implanted pedicle screws with varying backgrounds, screw head positions and changing illumination conditions. The images were captured by the main camera of the HoloLens at a resolution of 1344×756 pixels. The lumbosacral spine phantom in the images corresponds to one of the phantoms (phantom 1) used in the experiments. Another spine phantom with nerves (phantom 2) used for the experimental evaluation was not part of the training data set. A third spine phantom without nerves (phantom 3) was used for the evaluation of the trained network (see section 4). For training, all images were initially resized to 416×416 pixels. The network was trained without pre-trained weights for 100 000 iterations. The Adam optimizer was used to minimize the YOLO-loss. The learning rate was initially set to 0.0001, but scheduled to change later in training (1E-4 after 70 000 iterations and 1E-5 after 90 000 iterations). The training took 18 hours on our NVIDIA Quadro P6000.

The bounding box inference with HoloYolo is the first step of our pipeline. The inference works at interactive rates but it is slower than the rate of the forward-facing environmental cameras which operate at 30 frames per second. Employed on the Microsoft HoloLens, a single inference step took 900 ms. In order to maximize the flow of images through our pipeline, images were collected and fed into the network as soon as the previous inference ended.

3.2 Correspondence Finding

The output of Yolo are two sets of labelled bounding boxes with respective class confidences for the left and right camera images BB^L and BB^R, respectively. Finding correspondences from the n detected bounding box centers in the left image $c_i^L \in BB^L$ and the m detected bounding box centers in the right image $c_j^R \in BB^R$ is not straightforward, as the spatial relations between the centers are unknown and false detections can occur. The latter implies that n and m are not necessarily equal. We base our idea on correspondence finding on the observation that the two front-facing cameras are only slightly rotated against each other. By neglecting the rotational part, we can assume a pure translation between the cameras. This simplifies the problem of finding correspondence to finding a single translation in the 2D pixel space that is shared by a subset of all detected centers. In order to find this translation, connecting vectors $\boldsymbol{d}_k$ from every detected center in the left frame to every detected center in the right frame are formed (see Figure 2 C) resulting in $n \cdot m$ vectors $D = \{\boldsymbol{d}_k = c_i - c_j\} \forall\, c_i \in BB^L$ and $\forall\, c_j \in BB^R$. While most vectors will point in no consistent direction, vectors connecting correspondences will all point in a similar direction. Therefore, the problem of obtaining correspondence reduces to finding similar vectors which can be solved by transforming all vectors into a polar coordinate system as follows:

$$\alpha_k = atan2\left(\frac{d_{k,y}}{d_{k,x}}\right)$$

$$l_k = \sqrt{d_{k,x}^2 + d_{k,y}^2}$$

where α_k denotes the angle between the positive x-axis and the k[th] vector and l_k corresponds to its length. In this space, vectors are represented as points and similar vectors can be found by finding accumulations of points. We considered vectors within a radius of 10 units as similar (see Figure 2 D). If the minimum Euclidean distance between all possible point pairs was larger than this threshold, the entire image pair was rejected. Otherwise, the closest point pair was used as an accumulation seed, because it represents the most reasonable guess of a correct correspondence. Originating from the seed, all neighboring points lying within the radius of 10 units were included into the accumulation resulting in N correspondences. The corresponding centers (c_i^L, c_j^R) were decoded for each point of the accumulation set and used for the next step of the algorithm.

3.3 Triangulation

The corresponding image detections are individually extended by a third dimension (unit plane: z=1) such that they form 3D vectors $\boldsymbol{v}_k^L$ and $\boldsymbol{v}_k^R$ from their respective camera centers to the detection in their image, where k=1…N are the indices of the correspondences. The directional vectors were then extended to the rays $\boldsymbol{r}_k^L = \lambda_k^L \cdot \boldsymbol{v}_k^L$ and $\boldsymbol{r}_k^R = \lambda_k^R \cdot \boldsymbol{v}_k^R$, respectively. Both rays are expressed in their respective camera coordinate system. The known camera positions on the HoloLens in combination with the built-in HoloLens SLAM algorithm, that continuously estimates the head pose, allow to express the rays $\boldsymbol{r}_k^L$ and $\boldsymbol{r}_k^R$ in a global coordinate frame as $\boldsymbol{r'}_k^L$ and $\boldsymbol{r'}_k^R$. For each of the N correspondences, triangulation is performed by finding λ_k^L and λ_k^R, such that the distance between the two rays is minimized (See Figure 2 E).

$$\widehat{\lambda_k^L}, \widehat{\lambda_k^R} = \left\{ \underset{\lambda \in \mathbb{R}^+}{\operatorname{argmin}} \left(\boldsymbol{dist}(\boldsymbol{r'}_k^L(\lambda_k^L), \boldsymbol{r'}_k^R(\lambda_k^R)) \right) \right\} \; \forall \; k = 1, \ldots, N$$

Note that $\boldsymbol{dist}$ denotes the distance. The found $\widehat{\lambda_k^L}, \widehat{\lambda_k^R}$ define two points in 3D space which are fused into a single candidate point by linear interpolation.

Time synchronization was achieved by using the head pose provided by the HoloLens at the time of image acquisition for transforming the rays from their respective camera coordinate systems into a global coordinate frame. After triangulation, the 3D pedicle screw position estimates are still expressed in the same global coordinate frame which allows subsequent estimates from subsequent viewpoints to be merged by simply adding the new estimates to the existing set of candidate points.

3.4 Clustering

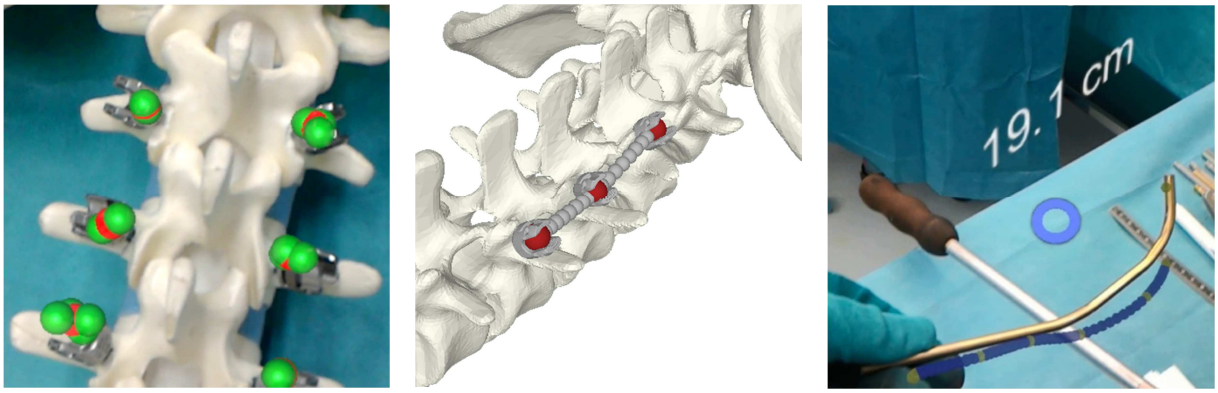
The previous step yields a set of 3D screw head candidates P_{cand} for every inference whereas not every screw is necessarily detected in every frame. However, this set may contain outliers. Therefore, an outlier removal procedure based on the key idea that all pedicle screws should be roughly co-planar to the coronal anterior-posterior (AP) plane is executed as follows:

Algorithm 1 Outlier Removal

$W_{up} \leftarrow (0,1,0)$
$P_{acc} \leftarrow \emptyset$
while collecting: **do**
  $P_{tot} \leftarrow P_{acc} \cup P_{cand}$
  $\mu \leftarrow mean(P_{tot} \cdot W_{up})$
  $\sigma \leftarrow stddev(P_{tot} \cdot W_{up})$
  $P_{acc} \leftarrow \emptyset$
  for all P_i in P_{tot} **do**
    if $P_i \cdot W_{up} \in [\mu - 1.5 \cdot \sigma, \mu + 1.5 \cdot \sigma]$ **then**
      $P_{acc} \leftarrow P_{acc} \cup P_i$

Initially, the HoloLens world up vector W_{up}, pointing upwards in the real world, is defined. Then, an empty set of accepted points P_{acc} is initialized. Afterwards, the following steps are repeated until the surgeon confirms completion of the procedure (see also Figure 2 F-H). Firstly, all previously accepted points P_{acc} are merged with the incoming new candidate points P_{cand} into a set P_{tot}. Secondly, mean and standard deviation of the points P_{tot} projected onto the world up vector W_{up} are determined and P_{acc} is reinitialized. Lastly, all points in P_{tot} whose projections $P_{tot} \cdot W_{up}$ lay within the interval $[\mu - 1.5 \cdot \sigma, \mu + 1.5 \cdot \sigma]$ are added to the accepted points P_{acc} while all other points are rejected.

The more correct screw head candidates are found, the narrower the acceptance interval will be. Additionally, since all points P_{tot} contribute to the mean and standard deviation, previously accepted screw candidates P_{acc} can fall out of consideration, if better points were collected in the meantime. The accepted points P_{acc} are visualized after each iteration to the surgeon such that he can stop the procedure as soon as the target precision is reached (see Figure 3).



(a) (b) (c)

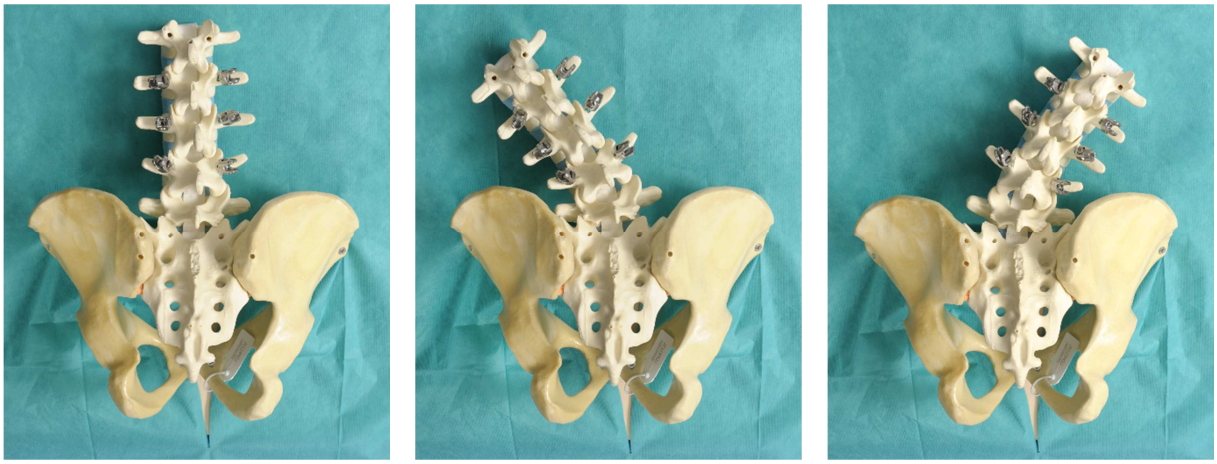
Figure 3: (a) Surgeon view: Detected screw head centers P_{acc} (green spheres) are constantly collected while the surgeon moves his head to capture different viewpoints. Once the surgeon is satisfied, the clustering is performed resulting in the clustered centers (red spheres). (b) The rod template is generated from the reconstructed 3D screw head positions. (c) The HoloLens displays the desired shape of the rod as well as its length to the surgeon.

Once all screws are detected, the surgeon has to confirm the correct result with a click gesture as an additional safety measure. Afterwards, a k-means++ clustering [32], initialized by the total number of screws, is performed to calculate the final screw positions. For navigating

the bending of the rod implant, the optimal shape and length are calculated from the screw positions [13]. Then, the rod is shortened to the desired length provided to the surgeon by the HoloLens. Next, the surgeon iteratively bends the rod and compares the intermediate rod to the target shape that is displayed as shown in Figure 3c. This approach allows the surgeon to bend the rod implant ex situ, reducing rebending maneuvers and surgery time [13] while simultaneously decreasing the risk of infection.

3.5 Validation setup and experiments

We compared the precision and duration of our navigation approach with the method suggested by Wanivenhaus et al. [13]. The validation setup consisted of two lumbosacral spine



(a) (b) (c)

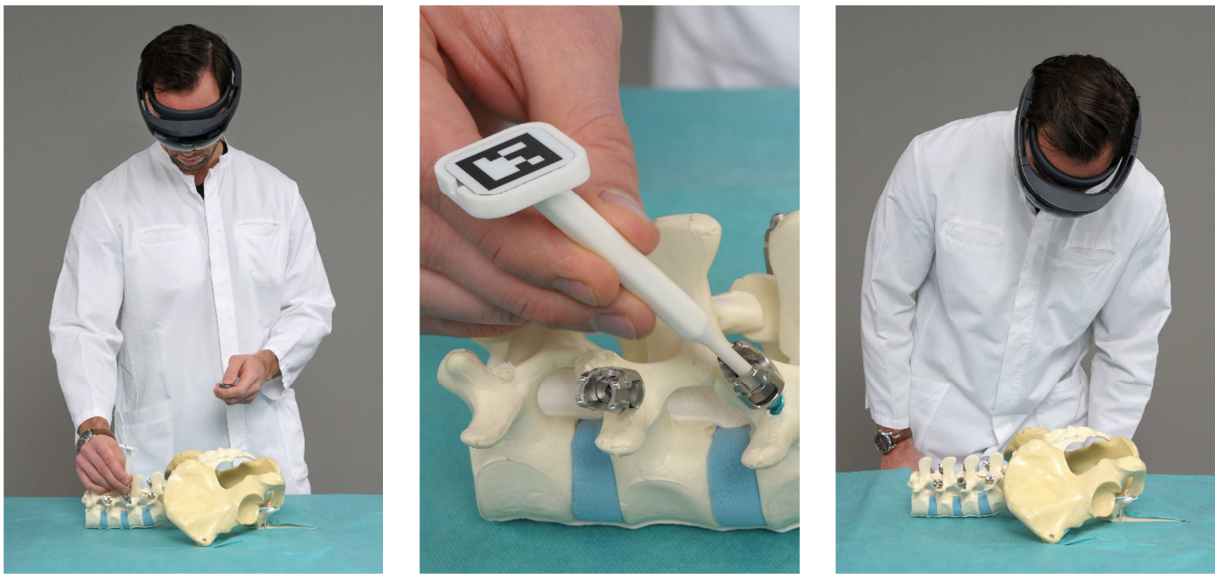
Figure 4: The three lumbar spine configurations shown on one of the spine phantoms. In addition to the physiological configuration (a) two pathological configurations where simulated in Figures (b) and (c).

phantoms (Synbone AG, Zizers, Switzerland) which were instrumented by pedicle screws (M.U.S.T., Medacta International SA, Switzerland) at three levels (L2--L4). The phantoms were fixated to a wooden board in three configurations each to simulate different deformities (see Figure 4). The phantoms differed in bone anatomy and the second phantom was more complex as it also contained additional structures such as nerves.

For each configuration, the following protocol was executed:

1. Computed Tomography (CT) scan according to clinical protocol [2]
2. Pointer-based acquisition of the screw head positions with a pointing device [13] (see Figure 5 A and B)
3. Marker-less acquisition using our proposed method (see Figure 5 C)

[2] 120 kV; 1 mm slice thickness; 0.5 mm slice increment; Somatom Edge Plus, Siemens, Munich, Germany



(a) (b) (c)

Figure 5: (a) Benchmark navigation method using a pointing device equipped with a trackable marker. (b) The surgeon acquires the screw position by placing the pointing device in the center of the screw head. (c) Our proposed method does not require a pointing device and the 3D positions of all screws are inferred solely from looking at the anatomy.

The experiments were conducted by two resident orthopedic surgeons with experience in spine surgery and surgical navigation. The navigation methods were applied successively and the experiments were repeated five times for each configuration. In each run, the estimated screw head positions were stored on the HoloLens and the required time was recorded.

CT's were acquired for each deformity configuration using a Siemens Somatom device (Siemens Healthineers, Erlangen, Germany) to obtain ground truth data for the accuracy evaluation. To this end, 3D triangular surface models of the screw heads were extracted from the CT data by using the thresholding and region growing functionality of a commercial segmentation software (Mimics version 19.0, Materialise NV, Leuven, Belgium). The original CAD models of the screw heads, in which the exact center points were indicated, were then registered with the segmented screw heads using the iterative closest point algorithm [33]. These centers then served as the ground truth screw head positions in the accuracy evaluation. The accuracy of the two methods was determined by comparing the stored positions against the CT-based ground truth. For this purpose, the three left and three right screw head positions of each run were registered in a least-squares sense [34] to the corresponding CT-based ground truth positions and the mean 3D distance errors were calculated. This resulted in a total of 240 data points (2 surgeons, 2 methods, 2 phantoms, 3 configurations, 5 runs, 2 sides: L/R).

3.6 Statistical evaluation

Equality to normal distribution was tested with the Kolmogorov-Smirnov test ($\alpha = 0.05$). Statistical differences with respect to precision and duration between the two methods were tested using a paired two sample t-test ($\alpha = 0.05$). Additionally, differences between the surgeons were examined using a two sample Kolmogorov-Smirnov test ($\alpha = 0.05$).

4 Results

The presented results are separated into an evaluation of our trained neural network and an assessment of the whole pipeline including the resulting pedicle screw head position estimates.

The trained network was evaluated against 200 unseen images of phantom 3 with varying background and illumination conditions and with twelve implanted pedicle screws. Half of the images in the test set were RGB images captured by the main camera of the HoloLens at a resolution of 1344×756 pixels and the other half were grayscale images taken by the front-facing environmental cameras of the HoloLens at a resolution of 480×640 pixels. The average precision (AP) of the model was evaluated against the RGB images and against the grayscale images separately at an IOU threshold of 0.25. The model achieved an AP of 69.34% when only evaluated on the grayscale images and an AP of 96.76% when considering RGB images.

Results of the accuracy evaluation are given in Figure 6. Our hands-free method achieved a mean distance error of 1.83±1.10 mm in the estimation of the 3D screw heads center compared to the CT ground truth. The minimum and maximum errors were 0.28 mm

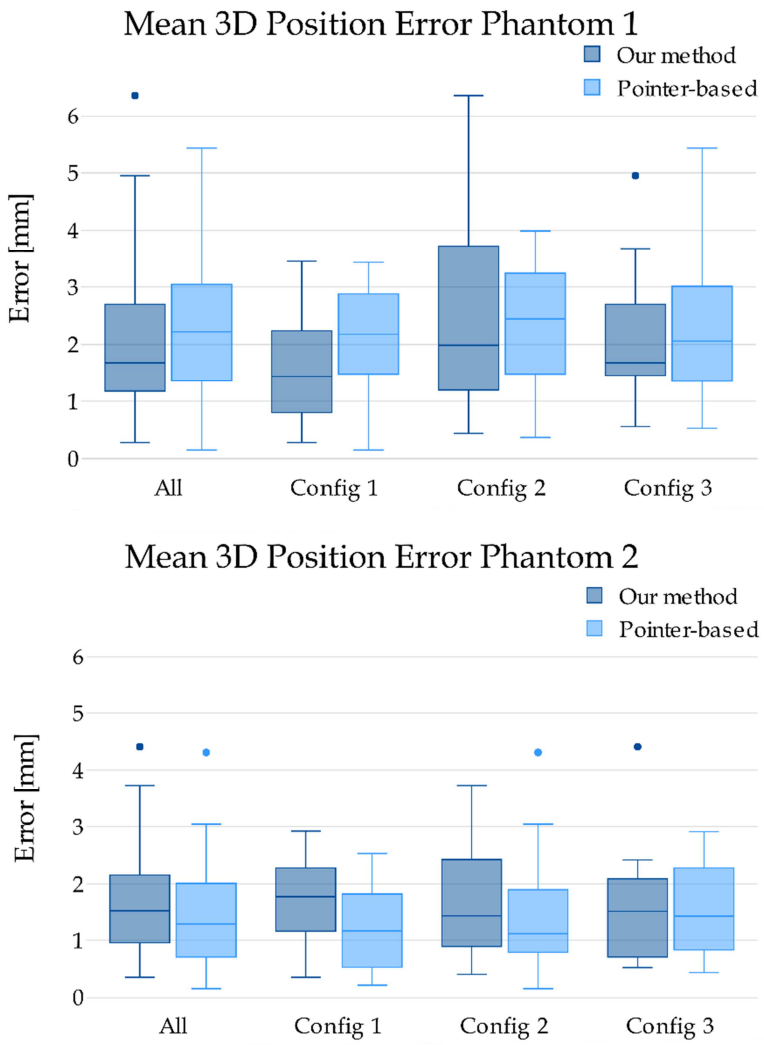
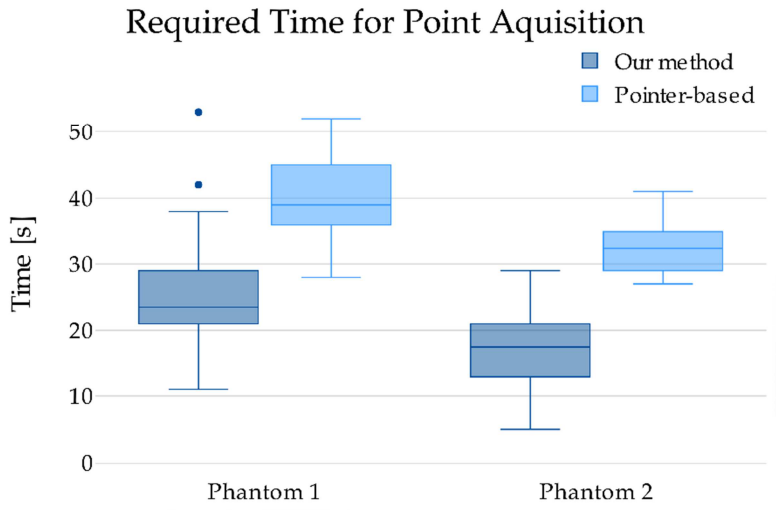


Figure 6: The error distribution of the mean accuracy in estimating the screw head center positions is depicted for each configuration and overall. Top: Phantom 1, Bottom: More complex phantom 2.

and 6.36 mm, respectively. The evaluation of the pointer-based benchmark navigation method resulted in a mean error of 1.87±1.31 mm. The error range of this method was between 0.15 mm and 8.19 mm.

On average, it took the surgeon 21.33±8.80 s to collect all 3D screw head centers using our hands-free method. The shortest run took 5 s and the longest lasted 53 s. To collect the six screw centers manually using the pointer-based method, the surgeon required 36.65±7.49 s, ranging from 27 s to 70 s, as can be observed in Figure 7. Note that we omitted three screws found by the pointer-based method, because the 3D position of the screw heads



was not successfully obtained.

The Kolmogorov-Smirnov test confirmed that all our data were distributed normally. The 3D distance error was not significantly different between the two methods ($p=0.08$). Contrary, a significant difference in the acquisition time ($p<0.01$) could be observed. Additionally, the two sample Kolmogorov-Smirnov test verified that the data of the two surgeons originated from the same distribution.

5 Discussion

Tracking of surgical tools and pointing devices by the means of optical markers is the gold standard in surgical navigation of orthopedic surgeries [35]. In recent years, augmented reality has become a mature technology offering new and exciting possibilities in the field of surgical navigation [10,11,36,37]. With the rise of machine learning, surgical navigation is shifting

Figure 7: The required time for acquiring the screw head centers is given for both phantoms.

from traditional methods to intelligent approaches that consider the surgical context and develop an understanding of surgical procedures [38]. In this work, we employed machine learning to transform the simple but tedious task of manually acquiring screw head positions with a pointing device into a completely machine-driven procedure.

Our proposed method achieved a similar average localization error of the pedicle screw heads when compared to the pointer-based approach of Wanivenhaus et al. [13]. Their benchmark method has proven to be accurate enough to provide significant advantages in the challenging process of rod bending which promotes the potential of our proof-of-concept

for clinical use. In addition to the accuracy, the time benefit when using our method may increase with a higher number of screws as multiple simultaneous detections are feasible.

Although our method has significantly reduced the time required to calculate screw head centers, we still consider an average time of 20 seconds as being too long. We are striving for real-time acquisition in which a single glance at the instrumented anatomy is sufficient to derive the optimal rod implant. One reason for the long acquisition time is the low computational power of the HoloLens. Our method remains robust although only a few stereo frames can be processed given the current frame rate of our pipeline of roughly 1 fps. Streaming to a computational server has been proposed as an alternative approach for solving the performance bottleneck [39]. However, we believe that on-device machine learning has unique benefits to streaming solutions such as immediate responses, enhanced reliability and increased privacy, particularly as portable device hardware becomes increasingly powerful. Another reason for the long acquisition time is the limited performance of our correspondence finding in the presence of only few or false detections. Other approaches to finding correspondence including the epipolar geometry were examined. For many viewpoints, however, the epipolar lines were too close to each other to assign detections to individual epipolar lines. This was mainly due to the co-linear arrangement of the implanted pedicle screws. On the other hand, by avoiding the epipolar geometry, our approach does not require exposure of internal information of the AR-hardware such as the camera intrinsics which not only makes our approach generic and device-independent, but also supersedes potential preparation work such as camera calibrations.

Our study has several limitations. Firstly, in our trials most of the six screw centers were detected after just a few seconds, but it took the algorithm much longer to identify the remaining screw centers. The variance in detection time may have resulted from an unfortunate combination of viewing angle and orientation of the screw head. The amount of training data in our network is still very limited and we believe that a significant increase in the detection rate can be achieved by including more labelled data.

Another limitation was the compromise of using the center of the regressed bounding box as the screw head center. Instead of taking the bounding box center, an alternative approach would be to regress the 2D pixel position of a screw head center directly using a different network architecture such as center voting (e.g. PoseCNN [40]). Unfortunately, these architectures are very complex and do exceed the current hard- and software capabilities of the HoloLens. Another hardware related constraint is that the two front-facing environmental cameras have a very limited field of view. Detection of a screw in both cameras is hence strongly dependent on the distance between camera and screw. If the distance is too short, the screw cannot be seen by both cameras and, vice-versa, the detection accuracy diminishes if the distance is too big. For this reason, our method works specifically well at a distance of approximately 30-50 cm, which is less than the suggested minimum working distance of 50 cm proposed by Microsoft. However, the visualization serves as an additional verification step to avoid mistakes during surgery and does not contribute to the accuracy of detected screws.

A further limitation is the validation setup that consists only of synthetic bones although we included several pathological configurations. Hence, our method has not been evaluated for large 3D deformities. However, we expect our outlier removal procedure to be sufficiently robust even in the case of larger deformities, because we introduced a rather

conservative acceptance interval for incoming screw position candidates. Lastly, despite the ease of use in surgery compared to a pointer, the HoloLens has shortcomings that might place a burden on surgeons such as wearing comfort due to the product weight.

As a next step, we will extend our method to work with real human anatomy by training HoloYolo with intraoperative data. This would enable us to assess performance in a real surgical setting. Additionally, we intend to include more object classes such as the wound, drills, clamps and other surgical tools. Furthermore, the resulting virtual model of the rod could be further exploited by 3D printers or robotic bending systems in order to fabricate the rod. These technologies, however, are not yet ready to manufacture rods within minutes, but could be interesting options in the future. Moreover, it would be interesting to explore the inclusion of other image modalities provided by the HoloLens to either refine the position estimations or to support the outlier detection.

In summary, our purely machine learning based proof-of-concept could achieve better or comparable accuracy than the benchmark navigation approach that require contact with the anatomy while requiring significantly less time to acquire the screw head positions compared to the marker-based benchmark method. In our opinion even more important is the demonstration on how the combination of new technologies can shape the surgery of the future.

Conflict of Interest

MF is shareholder and member of the board of directors of Incremed AG, a company developing mixed-reality applications. All other authors declare that they have no conflict of interest.

Compliance with Ethical Standards

– This article does not contain any studies with human participants or animals performed by any of the authors.

– This article does not contain patient data.

6 Bibliography

1. Kobayashi K, Ando K, Nishida Y, Ishiguro N, Imagama S. Epidemiological trends in spine surgery over 10 years in a multicenter database. *Eur Spine J*. 2018;27(8):1698-1703. doi:10.1007/s00586-018-5513-4
2. Martin BI, Mirza SK, Spina N, Spiker WR, Lawrence B, Brodke DS. Trends in Lumbar Fusion Procedure Rates and Associated Hospital Costs for Degenerative Spinal Diseases in the United States, 2004 to 2015. *Spine (Phila Pa 1976)*. 2019;44(5):369-376. doi:10.1097/BRS.0000000000002822
3. Tribus CB. Degenerative lumbar scoliosis: evaluation and management. *J Am Acad Orthop Surg*. 2003;11(3):174-183. doi:10.5435/00124635-200305000-00004
4. Mirza SK, Deyo RA. Systematic review of randomized trials comparing lumbar fusion surgery to nonoperative care for treatment of chronic back pain. *Spine (Phila Pa 1976)*. 2007;32(7):816-823. doi:10.1097/01.brs.0000259225.37454.38
5. Verlaan JJ, Diekerhof CH, Buskens E, et al. Surgical Treatment of Traumatic Fractures of the Thoracic and Lumbar Spine: A Systematic Review of the Literature on Techniques, Complications, and Outcome. *Spine (Phila Pa 1976)*. 2004;29(7):803-814. doi:10.1097/01.BRS.0000116990.31984.A9
6. Maruyama T, Takeshita K. Surgical treatment of scoliosis: A review of techniques currently applied. *Scoliosis*. 2008;3(1). doi:10.1186/1748-7161-3-6
7. Farshad M, Betz M, Farshad-Amacker NA, Moser M. Accuracy of patient-specific template-guided vs. free-hand fluoroscopically controlled pedicle screw placement in the thoracic and lumbar spine: a randomized cadaveric study. *Eur Spine J*. 2017;26(3):738-749. doi:10.1007/s00586-016-4728-5
8. Nottmeier EW, Crosby TL. Timing of paired points and surface matching registration in three-dimensional (3D) image-guided spinal surgery. *J Spinal Disord Tech*. 2007;20(4):268-270. doi:10.1097/01.bsd.0000211282.06519.ab
9. Richter M, Cakir B, Schmidt R. Cervical pedicle screws: Conventional versus computer-assisted placement of cannulated screws. *Spine (Phila Pa 1976)*. 2005;30(20):2280-2287. doi:10.1097/01.brs.0000182275.31425.cd
10. Liebmann F, Roner S, von Atzigen M, et al. Pedicle screw navigation using surface digitization on the Microsoft HoloLens. *Int J Comput Assist Radiol Surg*. 2019;14(7):1157-1165. doi:10.1007/s11548-019-01973-7
11. Elmi-Terander A, Burström G, Nachabe R, et al. Pedicle Screw Placement Using Augmented Reality Surgical Navigation With Intraoperative 3D Imaging: A First In-Human Prospective Cohort Study. *Spine (Phila Pa 1976)*. 2019;44(7):517-525. doi:10.1097/BRS.0000000000002876
12. Tohmeh A, Isaacs RE, Dooley ZA, Turner AW. Long Construct Pedicle Screw Reduction and Residual Forces are Decreased Using a Computer-Assisted Spinal Rod Bending System. *Spine J*. 2014;14(11):S143-S144. doi:10.1016/j.spinee.2014.08.348
13. Wanivenhaus F, Neuhaus C, Liebmann F, Roner S, Spirig JM, Farshad M. Augmented reality-assisted rod bending in spinal surgery. *Spine J*. 2019;19(10):1687-1689. doi:10.1016/j.spinee.2019.06.019
14. Ren S, He K, Girshick R, Sun J. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*.; 2016. http://image-net.org/challenges/LSVRC/2015/results. Accessed January 30, 2019.
15. Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. In: Vol 9905 LNCS. ; 2016:21-37. doi:10.1007/978-3-319-46448-0_2
16. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. 2015. doi:10.1109/CVPR.2016.91
17. Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Vol 2017-Janua. ;

2017:6517-6525. doi:10.1109/CVPR.2017.690
18. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. 2018. doi:10.1109/CVPR.2017.690
19. Petscharnig S, Schöffmann K. Learning laparoscopic video shot classification for gynecological surgery. *Multimed Tools Appl*. 2018;77:8061-8079. doi:10.1007/s11042-017-4699-5
20. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning. In: *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*. ; 2019:624-628. doi:10.1109/ICMLA.2018.00100
21. Pakhomov D, Premachandran V, Allan M, Azizian M, Navab N. *Deep Residual Learning for Instrument Segmentation in Robotic Surgery*. https://github.com/warmspringwinds/tf-image-segmentation. Accessed April 17, 2020.
22. Ni ZL, Bian G Bin, Xie XL, Hou ZG, Zhou XH, Zhou YJ. RASNet: Segmentation for Tracking Surgical Instruments in Surgical Videos Using Refined Attention Segmentation Network. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. ; 2019:5735-5738. doi:10.1109/EMBC.2019.8856495
23. Nwoye CI, Didier Mutter ·, Marescaux J, Padoy N. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. *Int J Comput Assist Radiol Surg*. 1958;14:1059-1067. doi:10.1007/s11548-019-01958-6
24. Du X, Kurmann T, Chang PL, et al. Articulated multi-instrument 2-d pose estimation using fully convolutional networks. *IEEE Trans Med Imaging*. 2018;37(5):1276-1287. doi:10.1109/TMI.2017.2787672
25. Kurmann T, Marquez Neila P, Du X, et al. *Simultaneous Recognition and Pose Estimation of Instruments in Minimally Invasive Surgery*.
26. Yoshimura M, Marinho MM, Harada K, Mitsuishi M. *Single-Shot Pose Estimation of Surgical Robot Instruments' Shafts from Monocular Endoscopic Images*.
27. Probst T, Maninis KK, Chhatkuli A, Ourak M, Poorten E Vander, Van Gool L. Automatic Tool Landmark Detection for Stereo Vision in Robot-Assisted Retinal Surgery. *IEEE Robot Autom Lett*. 2018;3(1):612-619. doi:10.1109/LRA.2017.2778020
28. Mourgues F, Devemay F, Coste-Manière E. 3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery. In: *Proceedings - IEEE and ACM International Symposium on Augmented Reality, ISAR 2001*. ; 2001:191-192. doi:10.1109/ISAR.2001.970537
29. Ye M, Johns E, Handa A, Zhang L, Pratt P, Yang GZ. Self-Supervised Siamese Learning on Stereo Image Pairs for Depth Estimation in Robotic Surgery. In: ; 2017:27-28. doi:10.31256/hsmr2017.14
30. Lu J, Jayakumari A, Richter F, Li Y, Yip MC. *SuPer Deep: A Surgical Perception Framework for Robotic Tissue Manipulation Using Deep Learning for Feature Extraction*. https://sites.google.com/ucsd.edu/super-framework/home. Accessed April 19, 2020.
31. Microsoft. API reference for Universal Windows Platform (UWP) apps. https://docs.microsoft.com/en-us/uwp/. Published 2019. Accessed March 3, 2019.
32. Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding. In: *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*. Vol 07-09-Janu. ; 2007:1027-1035.
33. Besl PJ, McKay ND. A Method for Registration of 3-D Shapes. *IEEE Trans Pattern Anal Mach Intell*. 1992;14(2):239-256. doi:10.1109/34.121791
34. Horn BKP, Hilden HM, Negahdaripour S. Closed-form solution of absolute orientation using orthonormal matrices. *J Opt Soc Am A*. 1988;5(7):1127. doi:10.1364/josaa.5.001127

35. Ewurum CH, Guo Y, Pagnha S, Feng Z, Luo X. Surgical navigation in orthopedics: Workflow and system review. In: *Advances in Experimental Medicine and Biology*. Vol 1093. Springer New York LLC; 2018:47-63. doi:10.1007/978-981-13-1396-7_4
36. Müller F, Roner S, Liebmann F, Spirig JM, Fürnstahl P, Farshad M. Augmented Reality Navigation for Spinal Pedicle Screw Instrumentation using Intraoperative 3D Imaging. *Spine J*. October 2019. doi:10.1016/j.spinee.2019.10.012
37. Tucker EW, Fotouhi J, Lee SC, et al. Towards clinical translation of augmented orthopedic surgery: from pre-op CT to intra-op x-ray via RGBD sensing. In: SPIE-Intl Soc Optical Eng; 2018:15. doi:10.1117/12.2293675
38. Vercauteren T, Unberath M, Padoy N, Navab N. CAI4CAI: The Rise of Contextual Artificial Intelligence in Computer-Assisted Interventions. *Proc IEEE*. 2019:1-17. doi:10.1109/JPROC.2019.2946993
39. Joachimczak M, Liu J, Ando H. Real-Time mixed-reality telepresence via 3D reconstruction with hololens and commodity depth sensors. In: *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Vol 2017-Janua. Association for Computing Machinery, Inc; 2017:514-515. doi:10.1145/3136755.3143031
40. Xiang Y, Schmidt T, Narayanan V, Fox D. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. 2017. doi:10.15607/RSS.2018.XIV.019

Acknowledgement
This work is part of ”SURGENT” under the umbrella of University Medicine Zurich/Hochschulmedizin Zürich, Switzerland. The NVIDIA Quadro P6000 used for this research was donated by the NVIDIA Corporation. We would like to thank the two participating surgeons Dr. med. Armando Hoch and Dr. med. David E. Bauer from the Balgrist University Hospital for their support in the experimental validation and their valuable clinical insights